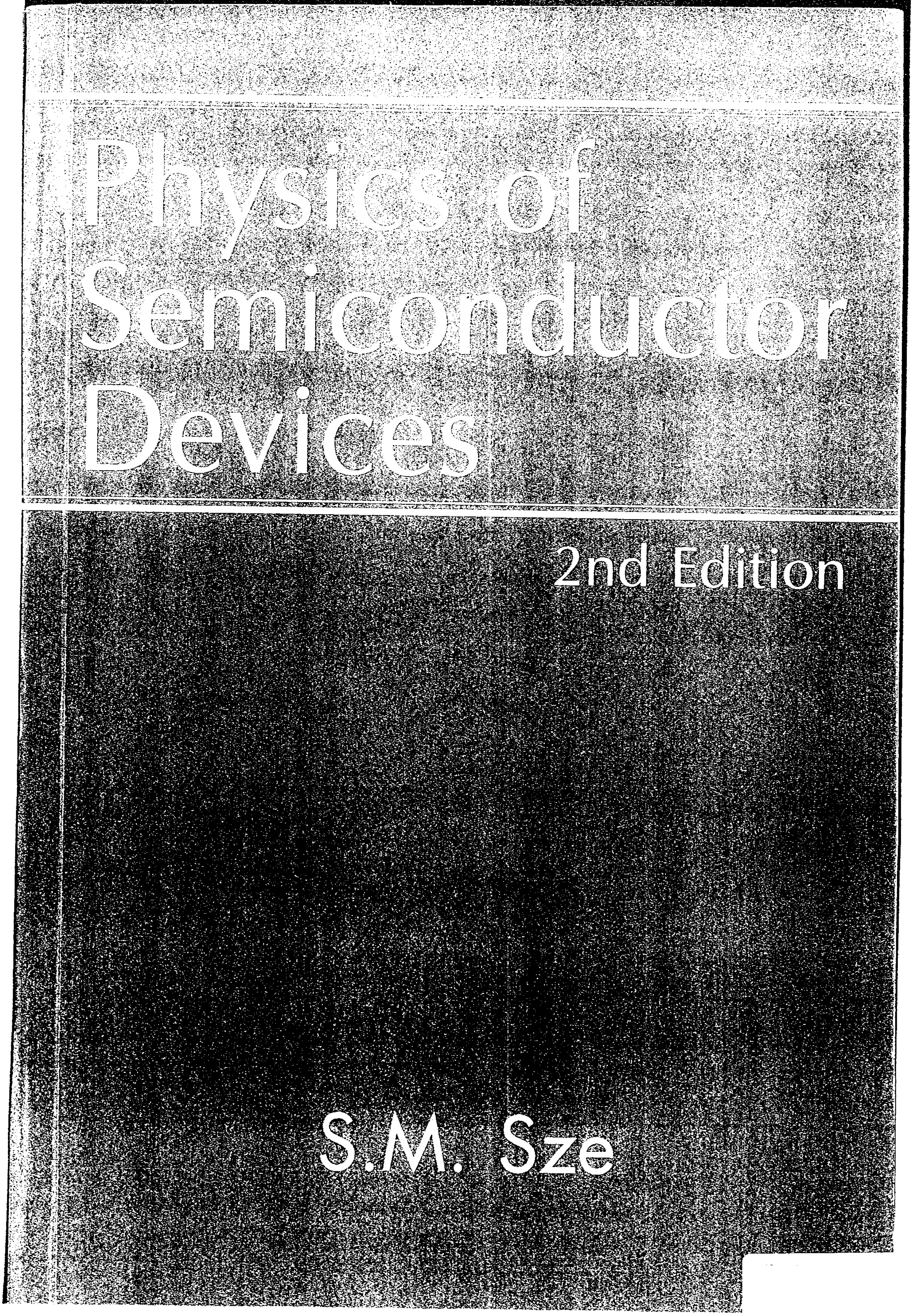
EXHIBIT E



Physics of Semiconductor Devices

2nd Edition

S.M. Sze

WSE

WILEY

John Wiley & Sons (ASIA) Pte Ltd
2 Clementi Loop #02-01
Singapore 129809

WILEY

ISBN 9971-51-266-1

8

MOSFET

8.1 INTRODUCTION

The metal–oxide–semiconductor field-effect transistor (MOSFET) is the most important device for very-large-scale integrated circuits such as microprocessors and semiconductor memories. MOSFET is also becoming an important power device. It has many acronyms including IGFET (insulated-gate field-effect transistor) MISFET (metal–insulator–semiconductor field-effect transistor) and MOST (metal–oxide–semiconductor transistor). The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld[1] and Heil.[2] It was subsequently studied by Shockley and Pearson[3] in the late 1940s. In 1960, Kahng and Atalla[4] proposed and fabricated the first MOSFET using a thermally oxidized silicon structure. The basic device characteristics have been subsequently studied by Ihantola and Moll,[5,6] Sah,[7] and Hofstein and Heiman.[8] The technology, application, and device physics have been reviewed by Wallmark and Johnson,[9] Richman,[10] and Brews.[11]

Because the current in a MOSFET is transported predominantly by carriers of one polarity only (e.g., electrons in an n-channel device), the MOSFET is usually referred to as a unipolar device. The MOSFET is a member of the family of field-effect transistors. The other members, JFETs and MESFETs, have already been considered in Chapter 6. Although MOSFETs have been made with various semiconductors such as Ge,[12] Si, and GaAs,[13] and use various insulators such as SiO_2, Si_3N_4, and Al_2O_3, the most important system is the Si–SiO_2 combination. Hence most of the results in this chapter are obtained from the Si–SiO_2 system.

We first consider the basic device characteristics of the so-called long-channel MOSFET; that is, the channel length L is much longer than the sum of the source and drain depletion-layer widths $(W_S + W_D)$.* This serves as a foundation to understand short-channel, that is, $L \lesssim (W_S + W_D)$, and related MOSFET devices.

Figure 1 shows[14] the reduction of the minimum device dimension since the beginning of the integrated circuit era in 1959. Figure 1 also shows that the minimum dimension will shrink continuously; the 1-μm barrier for commercial devices may be overcome by 1990. The reduction of device dimensions is driven by the requirement that integrated circuits of high complexity be fabricated. The number of components per integrated-circuit

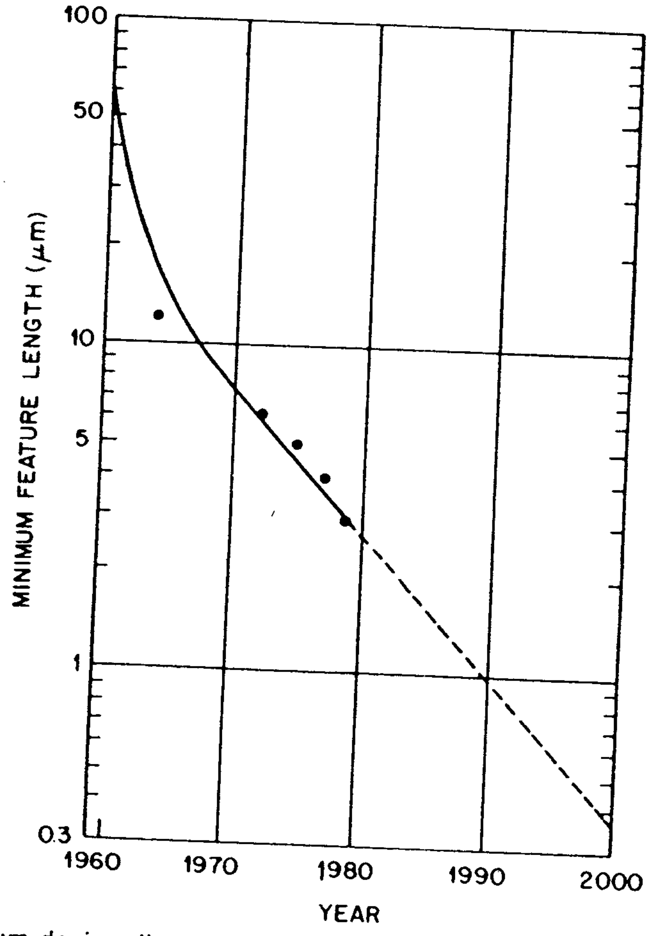


Fig. 1 The minimum device dimension in an integrated circuit as a function of the year for commercial devices. (After Ref. 14.)

*These terms will be defined in Section 8.2.

Basic D

Fig. 2

chip
expec
Howe
availa
chann
due t
Many
forma
condu
will b

8.2

Th
sistor
consi
the s
conta
comb
electi
the c
widtl
dopi
thick
isola
Th

*This
revers

s s' h as
i₃₁ and
nce most
n.
led long-
than the
).* This
$V_S + W_D$),

on since
ows that
rrier for
f device
of high
d-circuit

th ar

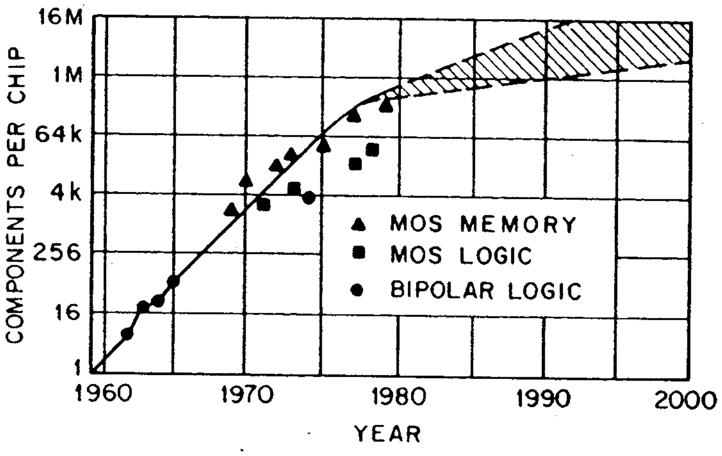


Fig. 2 Complexity of integrated circuits as a function of the year. (After Moore, Ref. 15.)

chip has grown exponentially[15] since 1959 (Fig. 2). The rate of growth is expected to slow down because of a lack of product definition and design. However, a complexity of 1 million or more devices per chip may be available around 1990 using 1-μm or submicron device geometries. As the channel length becomes shorter, one has to consider short-channel effects due to two-dimensional potential, high-field transport and oxide charging. Many device structures have been proposed to improve MOSFET performance. Some representative structures as well as the nonvolatile semiconductor memory, basically a MOSFET with a multilayer gate structure, will be discussed.

8.2 BASIC DEVICE CHARACTERISTICS

The basic structure of a metal–oxide–semiconductor field-effect transistor (MOSFET) is illustrated in Fig. 3. It is a four-terminal device and consists of a p-type semiconductor substrate into which two n^+ regions, the source and drain, are formed* (e.g., by ion implantation). The metal contact on the insulator is called gate; heavily doped polysilicon or a combination of silicide and polysilicon can also be used as the gate electrode. The basic device parameters are the channel length L, which is the distance between the two metallurgical n^+-p junctions; the channel width Z; the insulator thickness d; the junction depth r_j; and the substrate doping N_A. In a silicon integrated circuit, a MOSFET is surrounded by a thick oxide (called the field oxide to distinguish it from the gate oxide) to isolate it from adjacent devices.

The source contact will be used as the voltage reference throughout this

*This is an n-channel device; one may consider a p-channel device by exchanging p for n and reversing the polarity of the voltage.

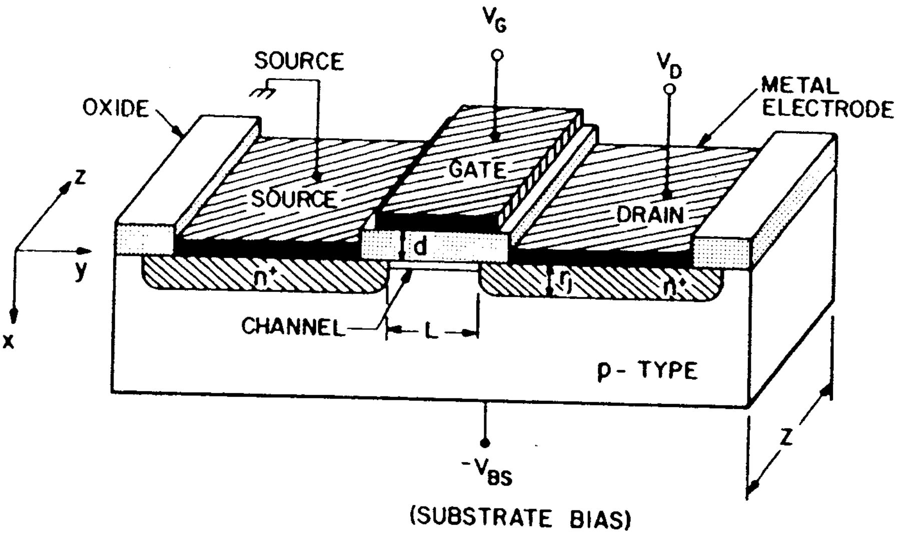


Fig. 3 Schematic diagram of a MOSFET. (After Kahng and Atalla, Ref. 4.)

chapter. When no voltage is applied to the gate, the source-to-drain electrodes correspond to two *p-n* junctions connected back to back. The only current that can flow from source to drain is the reverse leakage current.* When a sufficiently large positive bias is applied to the gate so that a surface inversion layer (or channel) is formed between the two n^+ regions, the source and the drain are then connected by a conducting-surface *n* channel through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltage. The back-surface contact (or substrate contact) can have the reference voltage or be reverse-biased; the back-surface voltage will also affect the channel conductance.

8.2.1 Nonequilibrium Condition

When a voltage is applied across the source–drain contacts, the MOS structure is in a honequilibrium condition; that is, the imref of the minority carriers (electrons, in the present case) is lowered from the equilibrium Fermi level. To show more clearly the band bending across the device, Fig. 4*a* shows[16] the MOSFET turned 90°. The two-dimensional, flat-band, zero-bias ($V_G = V_D = V_{BS} = 0$) equilibrium condition is shown in Fig. 4*b*. The equilibrium conditions under a gate bias that causes surface inversion are shown in Fig. 4*c*. The nonequilibrium condition with both drain and gate biases is shown in Fig. 4*d*, where we note the separation of the imrefs of electrons and holes; the hole imref E_{Fp} remains at the bulk Fermi level while the electron imref E_{Fn} (minority in the present case) is lowered

*This is the *n*-channel normally-off (enhancement-type) MOSFET. Other types will be discussed later.

Fig
tior
gat
Sat

tov
for
$\psi_s($
im
sui
]
bai
no
Ch

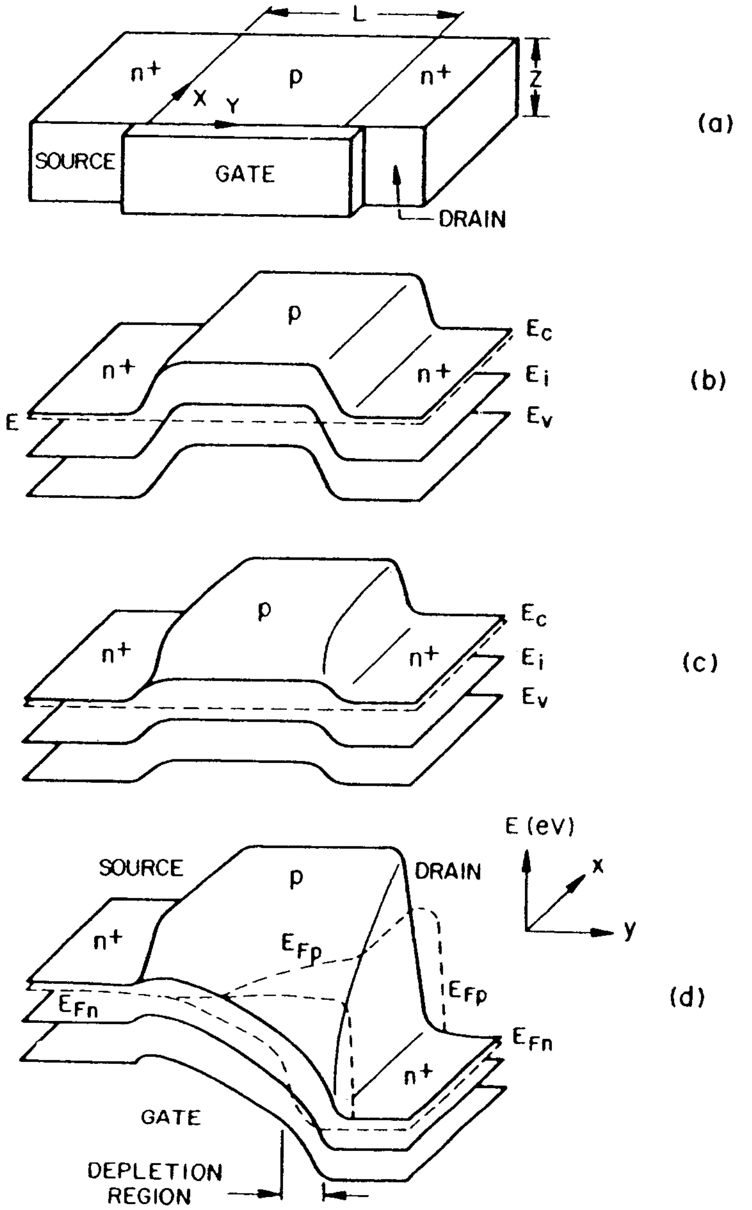


Fig. 4 Two-dimensional band diagram of an *n*-channel MOSFET. (*a*) Device configuration. (*b*) Flat-band zero-bias equilibrium condition. (*c*) Equilibrium condition under a gate bias. (*d*) Nonequilibrium condition under both gate and drain biases. (After Pao and Sah, Ref. 16.)

toward the drain contact. Figure 4*d* shows that the gate voltage required for inversion at the drain is larger than the equilibrium case in which $\psi_s(\text{inv}) \simeq 2\psi_B$. This is because the applied drain bias lowers the electron imref, and an inversion layer can be formed only when the potential at the surface crosses over the imref of the minority carrier.

Figure 5 shows a comparison[17] of the charge distribution and energy-band variation of an inverted *p* region for the equilibrium case and the nonequilibrium case at the drain. For the equilibrium case (discussed in Chapter 7), the surface depletion region reaches a maximum width W_m at

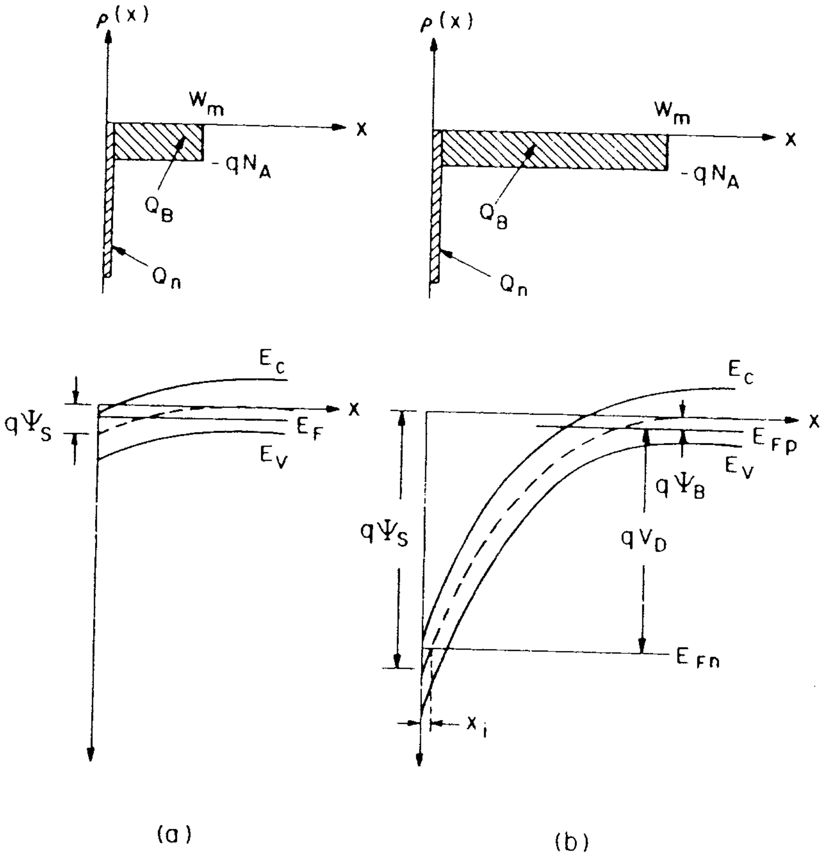


Fig. 5 Comparison of charge distribution and energy band variation of an inverted *p* region for (*a*) the equilibrium case and (*b*) the nonequilibrium case at the drain. (After Grove and Fitzgerald, Ref. 17.)

inversion. For the nonequilibrium case, the depletion-layer width is a function of the bias V_D, and the surface potential ψ_s at the onset of strong inversion is given, to a good approximation, by

$$\psi_s(\text{inv}) \simeq V_D + 2\psi_B . \qquad (1)$$

The derivation for the characteristic of the surface-space charge under the nonequilibrium condition is similar to that in Chapter 7. The two assumptions are that (1) the imref for the majority carriers of the substrate does not vary with distance from the bulk to the surface, and (2) the imref for the minority carriers of the substrate is separated by the applied junction bias V_D from the imref for the majority carriers; that is, $E_{Fp} = E_{Fn} + qV_D$ for a p substrate. The first assumption introduces little error when the surface is inverted, because majority carriers are then only a negligible part of the surface space charge; the second assumption is correct under the inversion condition, because minority carriers are an important part of the surface-space-charge region when the surface is inverted.

Based on these assumptions, the one-dimensional Poisson equation for

Basic

the sı

wher

Follc

and

wheı

and

The

whe

and

or

whe
imr
x_i i
forı

OSFET

ted p
After

is a
rong

(1)

ider
two
rate
iref
lied
p =
ror
y a

an
is

for

the surface-space-charge region at the drain is given by

$$\frac{\partial^2\psi}{\partial x^2} = -\frac{q}{\epsilon_s}(N_D^+ - N_A^- + p - n) \tag{2}$$

where

$$N_D^+ - N_A^- = n_{po} - p_{po}, \qquad p_{po} \simeq N_A$$
$$p = p_{po}e^{-\beta\psi}$$
$$n = n_{po}e^{\beta\psi-\beta V_D}, \qquad \beta \equiv q/kT. \tag{3}$$

Following the same approach as in Chapter 7, we obtain

$$\mathscr{E} = -\frac{\partial\psi}{\partial x} = \pm\frac{\sqrt{2}kT}{qL_D}F\left(\beta\psi, V_D, \frac{n_{po}}{p_{po}}\right) \tag{4}$$

and

$$Q_s = -\epsilon_s\mathscr{E}_s = \mp\frac{\sqrt{2}\epsilon_s kT}{qL_D}F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right) \tag{5}$$

where

$$F\left(\beta\psi, V_D, \frac{n_{po}}{p_{po}}\right) \equiv \left[e^{-\beta\psi} + \beta\psi - 1 + \frac{n_{po}}{p_{po}}e^{-\beta V_D}(e^{\beta\psi} - \beta\psi e^{\beta V_D} - 1)\right]^{1/2} \tag{6}$$

and

$$L_D \equiv \left(\frac{kT\epsilon_s}{p_{po}q^2}\right)^{1/2}. \tag{7}$$

The surface charge per unit area after strong inversion is given by

$$Q_s = Q_n + Q_B \tag{8}$$

where

$$Q_B = -qN_AW_m = -\sqrt{2qN_A\epsilon_s(V_D + 2\psi_B)} \tag{9}$$

and Q_n, the charge due to minority carriers within the inversion layer, is

$$|Q_n| \equiv q\int_0^{x_i} n(x)\,dx = q\int_{\psi_s}^{\psi_B}\frac{n(\psi)\,d\psi}{d\psi/dx} \tag{10}$$

or

$$|Q_n| = q\int_{\psi_s}^{\psi_B}\frac{n_{po}e^{(\beta\psi-\beta V_D)}\,d\psi}{(\sqrt{2}kT/qL_D)F(\beta\psi, V_D, n_{po}/p_{po})} \tag{11}$$

where x_i denotes the point at which the intrinsic Fermi level intersects the imref for electrons. For the practical doping ranges in silicon, the value of x_i is quite small, of the order of 30 to 300 Å. Equation 11 is the basic formula for long-channel MOSFET, and can be evaluated numerically.

Under strong inversion conditions, a simplified expression for Q_n can be obtained from a charge-sheet model[18] and is given by

$$|Q_n| = \sqrt{2} q N_A L_D \left\{ \left[\beta\psi_s + \left(\frac{n_{po}}{p_{po}}\right) e^{(\beta\psi_s - \beta V_D)} \right]^{1/2} - (\beta\psi_s)^{1/2} \right\}. \qquad (12)$$

This expression for Q_n is derived under the condition $V_{BS} = 0$. When a substrate reverse bias is applied, the depletion width increase, and the term βV_D in Eq. 12 is replaced by $\beta(V_D + V_{BS})$.

8.2.2 Linear and Saturation Regions

We shall first present a qualitative discussion of device operation. Let us consider that a voltage is applied to the gate, causing an inversion at the semiconductor surface, Fig. 6*a*. If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel. Thus the channel acts as a resistance, and the drain current I_D is proportional to the drain voltage V_D. This is the linear region. As the drain voltage increases, it eventually reaches a point at which the channel depth x_i at $y = L$ is reduced to zero; this is called the pinch-off point, Fig. 6*b*. Beyond the pinch-off point the drain current remains essentially the same, because for $V_D > V_{D\text{sat}}$, the voltage at Y remains the same, $V_{D\text{sat}}$. Thus the number of carriers arriving at point Y from the source, and hence the current flowing from source to drain, remains the same apart from a decrease in L to the value L' (Fig. 6*c*). Carrier injection from Y into the drain-depletion region is quite similar to the case of carrier injection from an emitter–base junction to the base–collector depletion region of a bipolar transistor.

We shall now derive the basic MOSFET characteristics under the following idealized conditions: (1) the gate structure corresponds to an ideal MOS diode as defined in Chapter 7; that is, there are no interface traps, fixed oxide charge, or work-function difference, and so on; (2) only drift current will be considered; (3) carrier mobility in the inversion layer is constant; (4) doping in the channel is uniform; (5) reverse leakage current is negligibly small; and (6) the transverse field ($\mathscr{E}_x$ in the x direction) in the channel is much larger than the longitudinal field ($\mathscr{E}_y$ in the y direction). The last condition corresponds to the so-called gradual channel approximation.

Under such idealized conditions, the total charge induced in the semiconductor per unit area Q_s at a distance y from the source is given by

$$Q_s(y) = [-V_G + \psi_s(y)]C_i \qquad (13)$$

where $C_i \equiv \epsilon_i/d$ is the capacitance per unit area. The charge in the inversion layer is given by

$$\begin{aligned} Q_n(y) &= Q_s(y) - Q_B(y) \\ &= -[V_G - \psi_s(y)]C_i - Q_B(y). \end{aligned} \qquad (14)$$

The surface potential $\psi_s(y)$ at inversion can be approximated by $2\psi_B +$

Ba

Fig
ope
ope

V(
ele
de|

Su

of device parameters. With other choices of device parameters, the relative importance of the various mechanisms changes.

To reduce the parasitic transistor effect, the resistance of the substrate R_{sub} can be minimized so that the product of the substrate current and R_{sub} remains smaller than 0.6 V when the drain voltage is equal to or larger than the corresponding BV_{CEO}. Then the breakdown voltage of a short-channel MOSFET will no longer be limited by BV_{CEO}; higher voltages and more reliable operation can be expected.[50] To reduce oxide charging, the density of water-related traps in the oxide should be minimized,[53] because such traps are known to capture electrons. To increase the punch-through voltage, single or double ion-implanted device structures can be made to increase the doping of the surface region. These structures will be considered in Section 8.5.

8.5 MOSFET STRUCTURES

Many device structures have been proposed to improve MOSFET performance with higher response speed, lower power consumption, more reliable operation, and higher power-handling capability. We shall now consider some representative structures.

8.5.1 Scaled-Down Device

In Section 8.4 we pointed out that short-channel effects are generally undesirable. One approach to avoid these effects is to maintain the long-channel behavior by simply scaling down all dimensions and voltages of a long-channel MOSFET, so that the internal electric fields are the same. This approach offers a conceptually simple picture for device miniaturization.

Figure 51*a* and *b* show the traditional large device and the scaled-down device,[54] respectively, in which all dimensions are shrunk by a "scaling factor," κ. This shrinking includes oxide thickness, channel length, channel width, and junction depth. The doping level is increased by κ, and all voltages are reduced by κ, leading to a reduction of the junction depletion width by about κ. Figure 51*c* compares I_D versus V_G in the linear region for the large and the scaled-down device. The threshold voltage is also reduced approximately by κ. Therefore, the number of devices per unit area increases by a factor of κ^2, the delay time due to transit across the channel, Eq. 54, decreases by κ, and the power dissipated per cell decreases by κ^2.

Note that in Fig. 51*c* the subthreshold current remains essentially the same for both devices. It remains the same because the subthreshold swing S, which is proportional to $(1 + C_D/C_i)$, remains the same as both capacitances are scaled up by the same factor κ. In addition, the junction

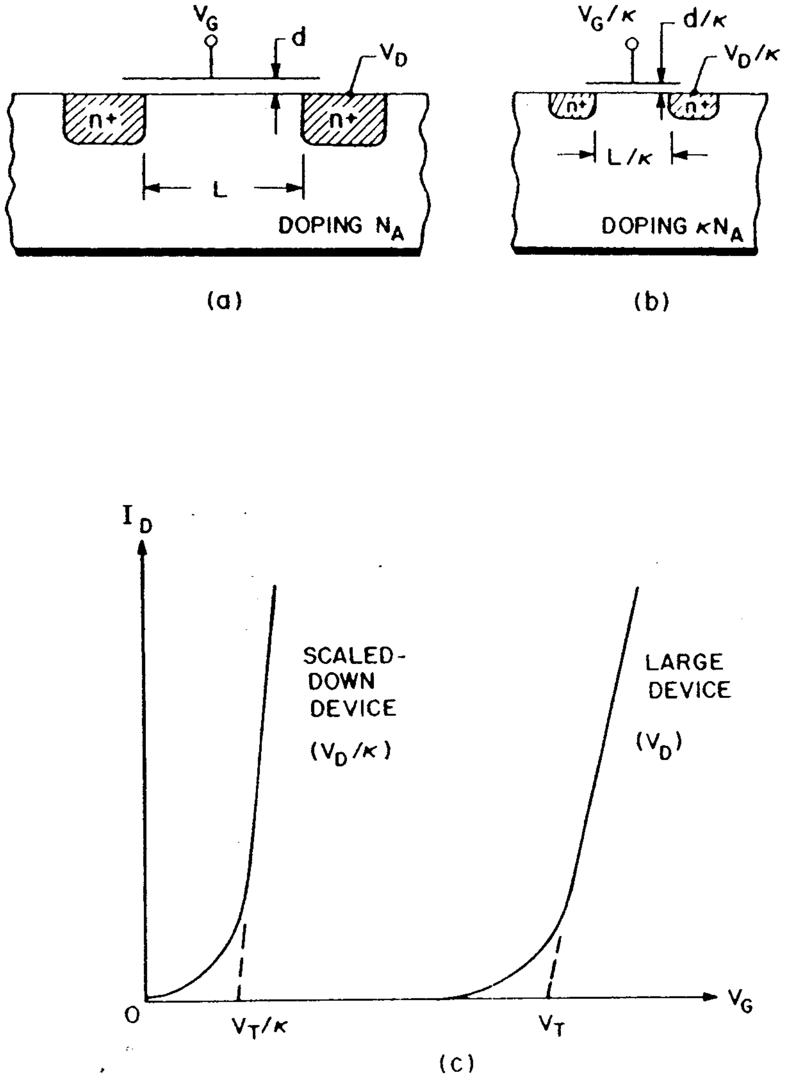


Fig. 51 Scaling approach for device miniaturization. (a) Long-channel device. (b) Scaled device. (c) Drain characteristics of these devices. (After Dennard et al., Ref. 54.)

built-in voltage and the surface potential for the onset of weak inversion do not scale (only ~10% change for 10 times increase in dopings). The range of gate voltage between depletion and heavy inversion is approximately 0.5 V. The parasitic capacitance may not scale, and the interconnect resistance increases when dimensions become smaller.

The expression for the minimum channel length, Eq. 83, can be used for a more flexible scaling approach.[41] For a given $L_{\min}$, the value of γ is obtainable from Eq. 83, or Fig. 35, which allows the various device parameters to be adjusted independently as long as the value of γ remains the same. Therefore, all device parameters do not have to be scaled by the same factor κ. This flexibility allows one to choose new geometries that are easier to make or which optimize other aspects of device operation, rather than choosing strictly scaled geometries.

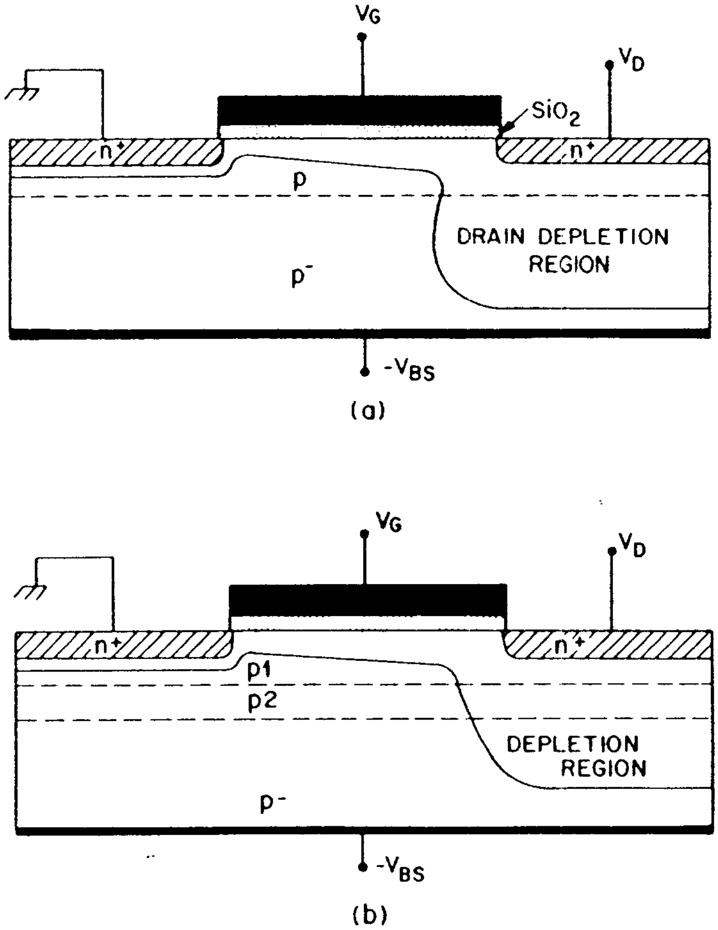


Fig. 52 HMOS structures. (a) Single implantation. (*b*) Double implantation. (After Shannon, Stephen, and Freeman, Ref. 55; Nihira et al., Ref. 56.)

8.5.2 HMOS

Figure 52 shows HMOS (high-performance MOS) structures. Figure 52*a* has a single ion implantation to increase the doping level at the surface region.[55] The implantation can control the threshold voltage and increase the punch-through voltage. Yet the surface region is shallow enough so that under operating conditions, the drain depletion-layer width extends into the low-doped substrate, reducing the drain capacitance. Figure 52*b* shows a double-implanted HMOS.[56] The $p1$ region contains the threshold control implant, and the $p2$ region contains the punch-through control implant. Using double implants, the HMOS with physical small-channel lengths can be tailored to minimize short-channel effects.

The implantations, however, degrade the subthreshold behavior[19] (large subthreshold swing) and can increase substrate bias sensitivity (becoming more sensitive to V_{BS}). However, various trade-offs exist and should be considered for device optimization.

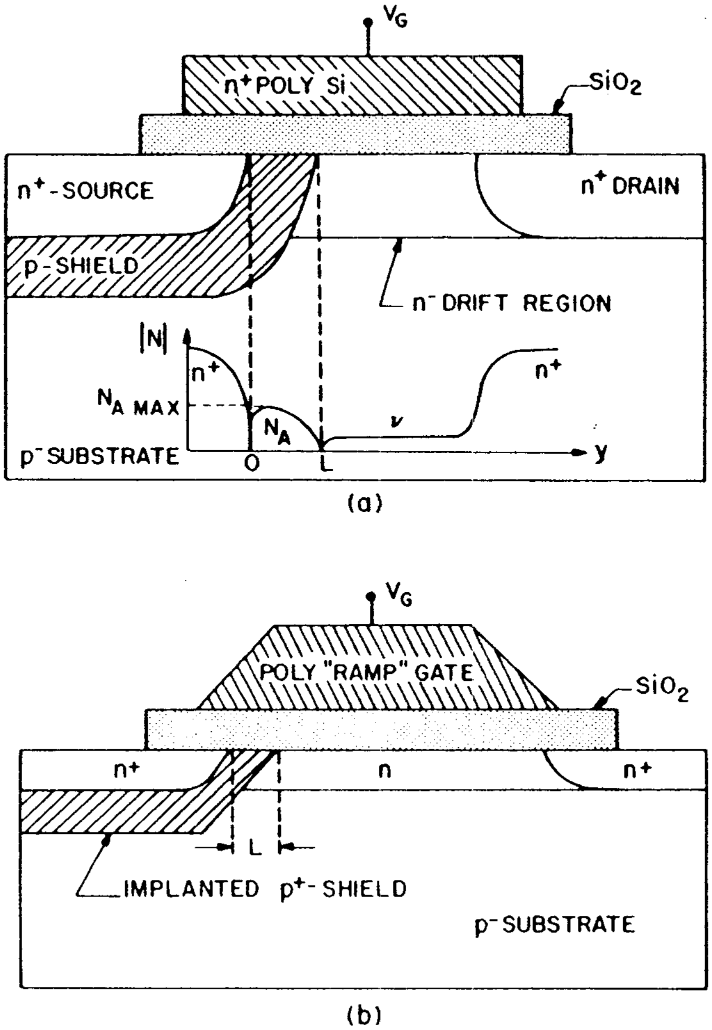


Fig. 53 (*a*) DMOS. (*b*) DIMOS structure. (After Tarui, Hayashi, and Sekigawa, Ref. 57; Tihanyi and Widmann, Ref. 58.)

8.5.3 DMOS

Figure 53*a* shows the DMOS (double-diffused MOS) structure,[57] where the channel length *L* is determined by the higher rate of diffusion of the *p*-dopant (e.g., boron), compared to the n^+-dopant (e.g., phosphorus) of the source. The channel is followed by a lightly doped drift region. Figure 53*a* also shows the doping profile along the semiconductor surface. Another version of DMOS is made by implantation. DIMOS (double-implanted MOS)[58] forms its source and drain by using a polysilicon gate as mask. The gate is tapered and the p^+-shield region is shaped by implantation through the tapered gate. The DIMOS structure improves the control in DMOS structures.

The DMOS and DIMOS structures can have very short channels and do not depend on a lithographic mask to determine channel length. Both structures have good punch-through control because of the heavily doped p-shield. The lightly doped drift region minimizes the voltage drop across the region by maintaining a uniform field ($\gtrsim 10^4$ V/cm) to achieve velocity saturation.[59] The field near the drain is the same as in the drift region, so avalanche breakdown, multiplication, and oxide charging are lessened, compared to conventional MOSFETs and HMOSs.[11]

However, the threshold voltage V_T is more difficult to control in DMOS.[60] As shown in Fig. 53*a*, V_T is determined by the maximum doping concentration $N_{A\,\max}$ along the semiconductor surface. Varying $N_{A\,\max}$ leads to variations in V_T. The localization of punch-through control to a thin p^+-shield region requires a higher doping level compared to HMOS, which leads to poorer turn-off behavior for DMOS.

8.5.4 Recessed-Channel MOSFET

The insert of Fig. 54 shows a MOSFET with a recessed channel.[61] The junction depth r_j for this structure is zero or negative. Figure 35 of Section 8.4 showed that the minimum channel length decreases as $r_j^{1/3}$. Figure 54 demonstrates that by reducing r_j, short-channel effects are minimized. For a given oxide thickness and substrate doping, as r_j decreases the onset of a large drop of V_T occurs at progressively shorter channels.

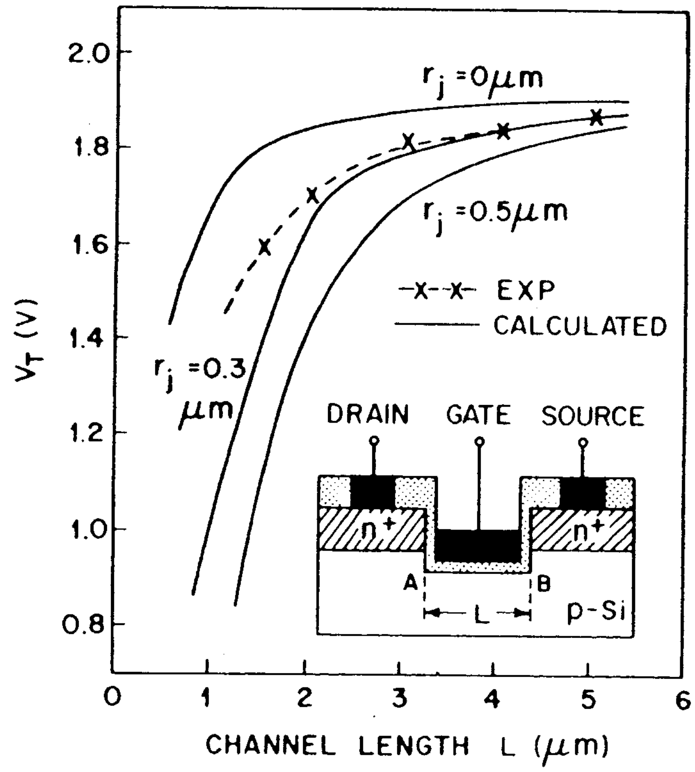


Fig. 54 Calculated and experimental V_T versus L plot for various junction depths. Insert shows a recessed-channel MOSFET. (After Nishimatsu et al., Ref. 61.)

Fig
dev
62.

mi
th
Al
in

re
sc
F
m
ca

ir

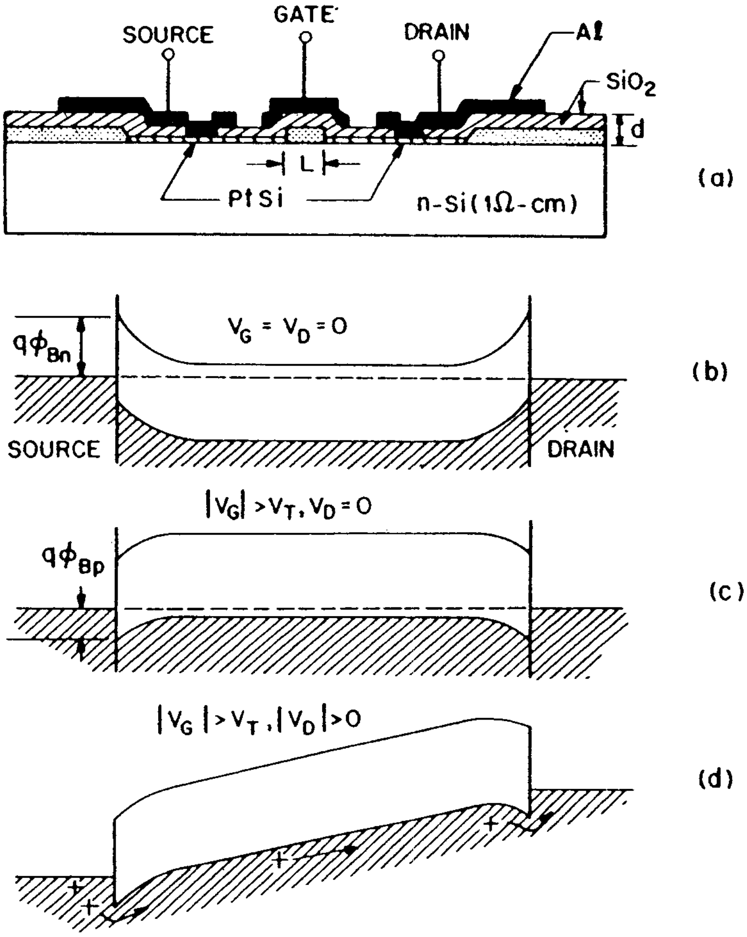


Fig. 55 MOSFET with Schottky-barrier source and drain. (a) Cross-sectional view of the device. (*b*), (*c*), and (*d*) Band diagrams for various biases. (After Lepselter and Sze, Ref. 62.)

The drawback of the recessed-channel structure, especially for submicron devices, is the difficulty in controlling the contour and the oxide thickness at corners A and B where the threshold voltage is determined. Also, oxide charging may be worsened, because more hot electron injection will occur.

8.5.5 Schottky-Barrier Source and Drain

Using Schottky-barrier contacts for the source and drain of a MOSFET results in performance and fabrication advantages. Figure 55*a* shows a schematic MOSFET structure with Schottky-barrier source and drain.[62] For a Schottky contact, the junction depth can effectively be made zero to minimize the short-channel effects. The high conductivity of the contact can also minimize source series resistance.

Eliminating high-temperature annealing steps can promote better quality in the oxides and better control of geometry. In addition, this structure can